DOCUMENT RESUME

ED 318 780 TM 014 884

AUTHOR Spearing, Debra; Woehlke, Paula
TITLE The Effect of Unequal Samples, Heterogeneity of Covariance Matrices, and Number of Variables on Discriminant Analysis Classification Tables and Related Statistics.
PUB DATE 89
NOTE 14p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classification; Correlation; *Discriminant Analysis; *Matrices; Monte Carlo Methods; *Sampling; Simulation; Tables (Data)
IDENTIFIERS *Covariance Matrices; *Heterogeneity of Variance

ABSTRACT

To assess the effect on discriminant analysis in terms of correct classification into two groups, the following parameters were systematically altered using Monte Carlo techniques: sample sizes; proportions of one group to the other; number of independent variables; and covariance matrices. The pairing of the off diagonals (or covariances) with the different numbers of variables and several different pairs of sample sizes within differing proportions required 396 Monte Carlo studies, each using 100 simulations. Results suggest that the proportion of one sample size to another is a major influence on correct classification. Correct classification was influenced by neither heterogeneity of covariance matrices, the number of variables used, nor a specific sample size when the proportional relationship between two samples was constant. The greatest hit rate occurred in the group with the largest sample. Statistics also found to be robust to heterogeneity of covariance matrices included: (1) eigenvalues; (2) canonical correlations; (3) Pillai's trace; and (4) Wilks' Lambda. Researchers using discriminant analysis to classify will be able to take into account and use the robustness of these statistics. Six bar graphs illustrate the study findings. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED318780

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DEBRA SPEARING

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Effect of Unequal Samples, Heterogeneity of Covariance Matrices, and Number of Variables on Discriminant Analysis Classification Tables and Related Statistics

Debra Spearing and Paula Woehlke

The (a) sample sizes, (b) proportions of one group to the other, (c) number of independent variables, and (d) covariance matrices were systematically altered using Monte Carlo techniques to assess the effect on discriminant analysis in terms of the correct classification into two groups.

Results from the 396 Monte Carlo studies used in the study suggest the proportion of one sample size to another is a major influence on correct classification. Correct classification was not influenced by heterogeneity of covariance matrices, the number of variables employed nor a specific sample size when the proportional relationship between two samples was constant. The greatest hit rate occurred in the group with the largest sample.

TM014984

BEST COPY AVAILABLE

When heterogeneity of variance is paired with unequal sample sizes in ANOVA, alpha may seriously be affected (Glass, Peckham, and Sanders, 1972). Olson (1976) and Stevens (1979) discovered excessively high Type I error rates for Roy's greatest root, Hotelling-Lawley trace, Wilks' Lambda, and Pillai-Bartlett trace when there was a violation of the MANOVA assumption of homogeneity of covariance matrices. MANOVA and discriminant analysis are mathematically the same with only the emphasis of canned computer programs differing. Tatsuoka (1988, pp. 351-358) Johnson and Wichern (1982, pp. 492-493), Klecka (1980, p. 61-62) and others assert heterogeneity of covariance matrices distorts the classification equations as well as the canonical discriminant functions. Prior to the present study, however, it was difficult to estimate how damaging the violation of the assumption of equal covariance matrices was to classification.

Classification procedures seek to discover a set of linear combinations for each group that will indicate the relative closeness, or "distance," of an individual case to each group centroid. The discriminating variables or the canonical discriminant functions are used to find the group to which the case most likely belongs. Overall adequacy of classification is not assessed with tests of significance but by the percent of correct classifications, or hit rates.

Johnson and Wichern (1982, p. 501) suggested it is very difficult to evaluate classification procedures analytically, but proposed generating different covariance structures on a computer. Monte Carlo studies have been used to study multivariate procedures such as factor analysis (Snook & Gorsuch, 1989; Seaman, 1986/1987; Shoemaker, 1982/1983). The present study used Monte Carlo techniques with

discriminant analysis to examine the effects of a number of characteristics, including heterogeneity of covariance matrices, on correct classification. More specifically, the following research question was evaluated by generating data from different covariance structures: As (a) sample sizes, (b) the proportions of one group to the other, (c) the number of independent variables vary while (d) the covariance matrices are systematically altered, what is the effect on the correct classification into the two groups?

Method

Data Generation and Collection

Multivariate standardized normal data were simulated around two covariance matrices with prespecified and uniform off diagonal values of .3, .5, or .7. Centroid differences were set to .5 standard deviations. A Fortran IMSL procedure generated pseudorandom scores while the SAS (1982) procedures of DISCRIM and CANDISC were used to produce classification tables and meaningful statistics. The DISCRIM procedure routinely used the proportional priors option. The numbers of independent variables used were three, five, and seven. The pairing of the off diagonals, or covariances, with the different number of variables and several different pairs of sample sizes within differing proportions required 396 Monte Carlo studies. Each Monte Carlo study used 100 simulations. Classification results from the SAS procedures were gathered and subjected to a program which produced measures of central tendency. Positive and negative hit rates were assessed with a

comparison to the known classification schemes as well as the percentage of correct classifications that may have been expected by chance.

Scores generated by the FORTRAN program were reviewed for skewness and the covariances were inspected for their conformation to the prespecified values that were submitted to the FORTRAN program. When sample sizes were equal and off diagonals were equal, the proportions of negative and positive hit rates were equal as was expected. As a final quality control measure, SPSSX (1986) was used to confirm the findings of the behavior of hit rates with the extreme ratio of 5:1.

Results

Hit Rates

The classification schemes from a discriminant analysis may be evaluated by comparing the hit rates from the analysis with hit rates which occur by chance. In the present study a positive hit rate refers to the percent of cases which are correctly classified into group one and a negative hit rate as the percent of cases which are correctly classified into group two. Positive (or negative) hit rates which occur by chance are calculated in two steps. The chance expected frequency for the respective cell is obtained by dividing the product of the row frequency and the column frequency for the cell by the total sample size. Chance hit rates are then computed by dividing the chance expected frequency of the cell by the respective row frequency. The number of variables and the covariances are not used in this calculation and would be identical across changes in these values but the expectation does change when the proportion of sample sizes changes.

When the sample sizes are in the proportion of a:b the method of obtaining hit rates which might be attributed to chance may be simplified. The positive and negative chance hit rates are calculated by dividing a by a plus b, and by dividing b by a plus b, respectively.

The obtained mean positive hit rates and chance hit rates for the different number of variables and covariances were graphed across the pairs of sample sizes. Negative hit rates were graphed in the same manner. Negative hit rates were identical to positive hit rates when proportions were reversed. For example, the negative hit rates with proportions of 5:1, 2:1, and 3:2 were identical to the positive hit rate with sample proportions of 1:5, 1:2, and 2:3, respectively. Hit rates were identical despite changes in the values of the off diagonals.

Positive Hits

Positive hit rates for the different number of variables and the positive hit rates by chance are illustrated in Figures 1a, 1b, and 1c, while negative hit rates are compared to chance in Figures 2a, 2b, and 2c. Covariances of .3 are used. Positive hit rates for the sample size ratios of 5:1, 2:1, and 3:2 are illustrated in Figure 1a and cannot be distinguished from negative hit rates for sample size ratios of 1:5, 1:2, and 2:3 (Figure 2b). The hit rates for the respective proportions and the different number of variables are the same to two digits. In the same manner, Figure 1b showing the positive hit rates for sample size ratios of 1:5, 1:2, and 2:3 is identical to the negative hit rates for sample size ratios of 5:1, 2:1, and 3:2 (Figure 2a). When the larger sample size of a proportion is paired with group one, the

positive hit rates are large and resemble the negative hit rates when the larger sample size of a proportion is paired with group two. The converse is also true. Positive hit rates for equal sample sizes are identical to the negative hit rates for equal sample sizes (Figure 1c and 2c).

Figure 1a reflects the positive hit rates of the different number of variables and the hit rates by chance across the proportions of 5:1, 2:1, and 3:2. The rate of correctly classifying cases from group one into group one was above .90 for ratios of 5:1 and 2:1, and above .84 for sample sizes with ratios of 3:2. The chance positive hit rates for these proportions are .8333, .6666, and .6000, respectively. The positive hit rates for the proportions of sample sizes of 1:5, 1:2, and 2:3 are shown in Figure 1b. The positive hit rates from the proportion of 1:5 are below .163 and are less or approximately equal to the positive hit rate which would occur by chance. The positive hit rates associated with the sample proportions of 1:2 and 2:3 fell between .1877 and .4820 and appear to approximate the values which may occur by chance. The chance rates for these two proportions are .3333 and .4000, respectively. The sample sizes are equal in Figure 1c and the positive hit rates from the Monte Carlo studies are above .6 while the chance hit rate is .5000.

As the proportions of sample sizes varied, the hit rates varied, but as the sample sizes varied or the number of variables changed while the proportion of one sample size remained constant, there was no appreciable change in the hit rates. It is the proportion of one sample

to the other that influences the hit rates rather than the size of the samples or number of variables used.

Hit Rates Using Different Covariances

Results for matrices with covariances other than .3 and .3 did not differ appreciably from the results discussed in the previous sections. A closer look at differences between pairs of covariances, however, will be discussed before leaving the topic of hit rates.

The twelve pairs of sample sizes from the proportions of 1:5 and 5:1 exhibit the more extreme hit rates and were chosen for a more detailed examination across the six different combinations of covariances. The covariances were: .3 and .3; .3 and .5; .3 and .7; .5 and .5; .7 and .5; and .7 and .7 for the two matrices, respectively. Positive and negative hit rates fell below .2 when the larger of the two sample sizes was paired with first group two and then group one respectively. Positive and negative hit rates fell above .9 when the larger of the sample sizes was paired with first group one and then group two respectively.

Conclusions

Findings from the current study with regard to correct classifications encompass the effects of different values of sample sizes, different proportions of one group to the other, the number of independent variables, and different values of off diagonals of covariance matrices. Different proportions of one group to the other influenced positive and negative hit rates but the other variables did not. When a five-to-one and a one-to-five ratio were generated,

positive and negative hit rates, respectively, were above .90 and decreased to slightly above .60 as the proportions became equal. Some statistical programs have the option of specifying proportional prior probabilities; however, such probabilities refer to the probabilities that exist in the population. Classification tables with and without the option of proportional priors from Monte Carlo studies were compared. Correct classifications were greater when the proportional option was included, as expected.

Eigenvalues, canonical correlations, Pillai's trace and Wilks' Lambda, in addition to hit rates, were also examined. These statistics were found to be robust to heterogeneity of covariance matrices, different sample sizes and the differing numbers of variables.

This study may be used by applied researchers as a demonstration of the serious effects of using disproportional sample sizes on classification schemes as well as the ineffectiveness of increasing sample size without equalizing proportions. Researchers using discriminant analysis to classify will now be able to take into account and use to their advantage the robustness to heterogeneity of covariance matrices, different sample sizes, and number of variables with the classification function of discriminant analysis, as well as eigenvalues, canonical correlations, Pillai's trace, and Wilks' Lambda.

REFERENCES

Glass, G. V., Peckham, P. D. & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

IMSL Library (1988). (Version 10.0). Houston, TX: Fortran subroutines for statistical analysis.

Johnson, N. & Wichern, D. (1982). Applied multivariate statistical analysis. Englewood Cliffs, New Jersey: Prentice-Hall.

Klecka, W. R. (1988). Discriminant analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-019. Beverly Hills and London: Sage Publications.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 83, 579-586.

Seaman, M. H. (1987). Dichotomous variables in factor analysis: A monte carlo simulation study. (Doctoral dissertation. Southern Illinois University, Carbondale, Illinois, 1986). Dissertation Abstracts International, 48, 03A.

Shoemaker, T. L. (1983). A monte carlo study of the problem of sample size in factor analysis (Doctoral dissertation. Southern Illinois University, Carbondale, Illinois, 1982). Dissertation Abstracts International, 44, 01B.

Snook, S. C. & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A monte carlo study. Psychological Bulletin, 106, 148-154.

Statistical Analysis System (1982). Cary, N.C.: SAS Institute Inc.

Statistical Package for the Social Sciences (1986). Chicago, IL: SPSS Inc.

Stevens, J. (1979). Comment on Olson: Choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 86, 355-360.

Tatsuoka, M. M. & Lohnes. P. R. (1988). Multivariate analysis (2nd ed.). New York: Macmillan.

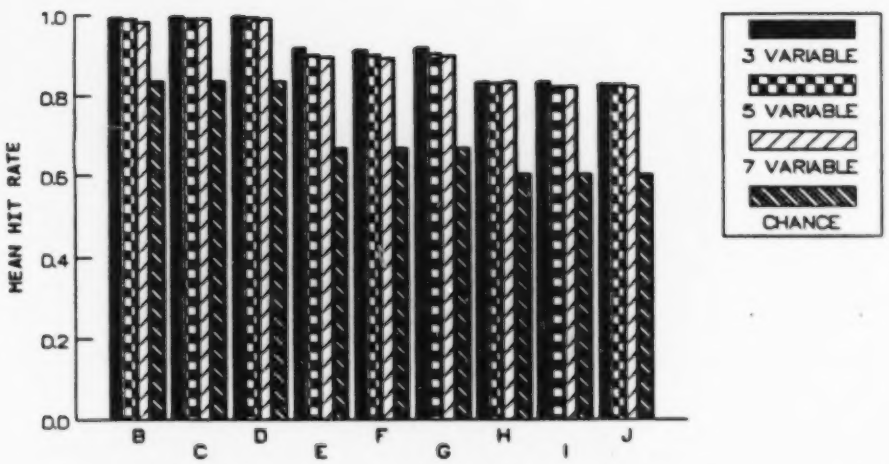

FIGURE 1A. MEAN POSITIVE HIT RATE FOR SAMPLE SIZE RATIOS OF 5:1, 2:1, AND 3:2 WITH OFF DIAGONALS OF .3 AND .3 FOR 3, 5, AND 7 VARIABLES AND CHANCE.

B 100:20	H 60:40	N 25:50	T 25:25	
C 300:60	I 99:66	O 50:100	U 50:50	
D 500:100	J 300:200	P 100:200	V 100:100	
E 50:25	K 20:100	Q 40:60	W 200:200	
F 100:50	L 60:300	R 66:99		
G 200:100	M 100:500	S 200:300		

FIGURE 1B. MEAN POSITIVE HIT RATE FOR SAMPLE SIZE RATIOS OF 1:5, 1:2, AND 2:3 WITH OFF DIAGONALS OF .3 AND .3 FOR 3, 5, AND 7 VARIABLES AND CHANCE.

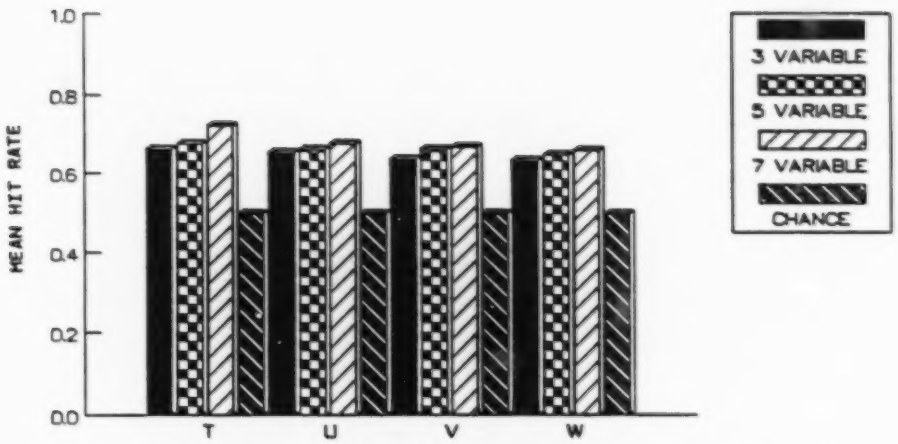

FIGURE 1C. MEAN POSITIVE HIT RATE FOR SAMPLE SIZE RATIOS OF 1:1 WITH OFF DIAGONALS OF .3 AND .3 FOR 3, 5, AND 7 VARIABLES AND CHANCE.

B 100:20	H 60:40	N 25:50	T 25:25	
C 300:60	I 99:66	O 50:100	U 50:50	
D 500:100	J 300:200	P 100:200	V 100:100	
E 50:25	K 20:100	Q 40:60	W 200:200	
F 100:50	L 60:300	R 66:99		
G 200:100	M 100:500	S 200:300		

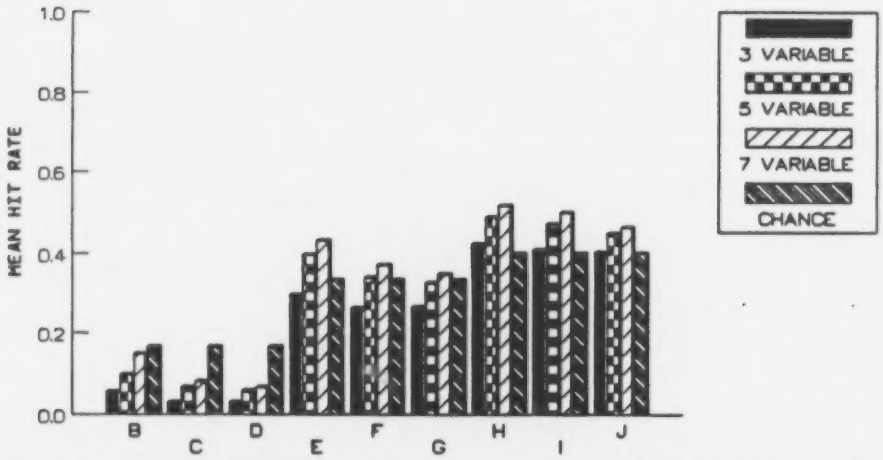

FIGURE 2A. MEAN NEGATIVE HIT RATE FOR SAMPLE SIZE RATIOS OF 5:1, 2:1, AND 3:2 WITH OFF DIAGONALS OF .3 AND .3 FOR 3, 5, AND 7 VARIABLES AND CHANCE.

B 100:20	H 60:40	N 25:50	T 25:25	
C 300:60	I 99:66	O 50:100	U 50:50	
D 500:100	J 300:200	P 100:200	V 100:100	
E 50:25	K 20:100	Q 40:60	M 200:200	
F 100:50	L 60:300	R 66:99		
G 200:100	M 100:500	S 200:300		

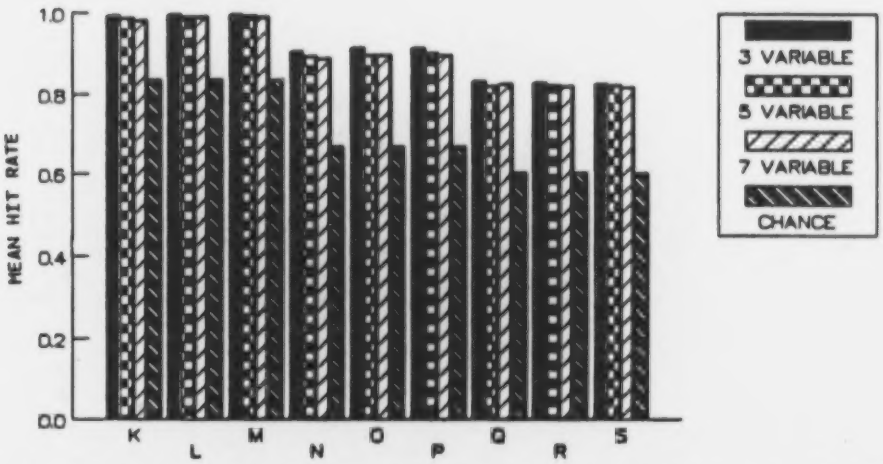

FIGURE 2B. MEAN NEGATIVE HIT RATE FOR SAMPLE SIZE RATIOS OF 1:5, 1:2, AND 2:3 WITH OFF DIAGONALS OF .3 AND .3 FOR 3, 5, AND 7 VARIABLES AND CHANCE.

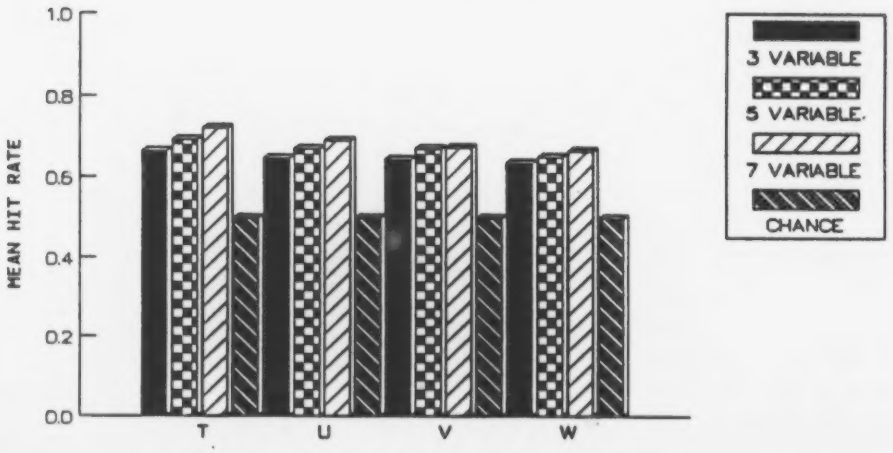

FIGURE 2C. MEAN NEGATIVE HIT RATE FOR SAMPLE SIZE RATIOS OF 1:1 WITH OFF DIAGONALS OF .3 AND .3 FOR 3, 5, AND 7 VARIABLES AND CHANCE.

B 100:20	H 60:40	N 25:50	T 25:25	
C 300:60	I 99:66	O 50:100	U 50:50	
D 500:100	J 300:200	P 100:200	V 100:100	
E 50:25	K 20:100	Q 40:60		M 200:200
F 100:50	L 60:300	R 66:99		
G 200:100	M 100:500	S 200:300		